STAT 300 Chapter 3

Zepu Zhang
February 27, 2012

1 Random variables

Definition A rule that associates a number with each outcome in the sample space $\mathcal{S}$ of a given experiment (Chap. 3.1). It is a function whose domain is $\mathcal{S}$ and whose range is $\mathcal{R}$.

It's a variable because there are more than one value it can assume. It's random because its value depends on the outcome, which is uncertain before the experiment is conducted.

This definition distinguishes the "outcome" of an experiment and the "number" associated with it. While the outcome may have clear descriptions, one has flexibilities in what rule to use to assign a number to it. For example, the experiment is to randomly pick a student ID number and record the gender of the student. The outcomes are "male" and "female". One may associate these outcomes with 0 and 1, respectively, or 1 and 2, or 2 and 1, or -1 and 1, etc.

This definition is the more formal way. In actual thinking we often jump over the "function" notion and think of each outcome as a number directly.

There are several situations in this function assignment. For example:
(1) Sample space is small, the number association is explicitly listed; Ex. 3.1 in Chap. 3.1.
(2) Number association is described in words; Ex. 3.2 in Chap. 3.1.
(3) Number association is defined in some more quantitative form;
(4) The outcome is the measurement of something (say temperature), and the variable is naturally the measured value; Ex. 3.5 in Chap. 3.1.
(5) One outcome may be associated with more than one number, i.e. a vector; then the random variable may be multivariate; Ex. 3.3 in Chap. 3.1.

(6) Infinitely many possible values: Ex. 3.4 in Chap. 3.1.

Discrete vs continuous random variables

Discrete random variable: finite or countably infinite possible values.

Continuous random variable: range is all numbers in an interval on $\mathcal{R}$ or union of intervals.

Example Bernoulli random variable: binary.

Notation Use upper case (X, Y, Z, ...) for a random variable and the corresponding lower case (x, y, z, ...) for a particular value of the variable. For example, $P(X = x)$ means the probability that the random variable X takes the particular value x. Further, suppose we know

$$P(X = x) = 0.3^x\, 0.7^{1-x}, \quad x = 0, 1$$

Then we know

$$P(X = 0) = 0.3^0\, 0.7^{1-0} = 0.7$$

and

$$P(X = 1) = 0.3^1\, 0.7^{1-1} = 0.3$$

Note that we should not write $P(x = 0) = 0.7$ or $P(x = 1) = 0.3$. Strictly speaking, the notation $P(x = 0)$ is meaningless, because x is not (or should not be used to denote) a random variable.

2 Probability mass function (pmf) of a discrete distribution

For a discrete random variable X, $p(x)$ is the probability that X takes value x, or, more rigorously, probability of the event that consists of the outcome(s) that is associated with the value x. If multiple outcomes are associated with the same value x, then $p(x)$ is the probability of the event defined by those outcomes.

A function that specifies the value $p(x)$ for every possible value x of the discrete variable X is called the probability mass function. (Sometimes also called the “probability distribution function”, but pmf is better).

This function describes how the total probability 1 is allocated to the various values of X. Value x gets probability mass $p(x)$. This may be an exhaustive listing (when X can take only a few values) or a formula.

Conditions (or properties):
(1) $p(x) \geq 0$;
(2) $\sum_{\text{all possible } x} p(x) = 1$.

Example Ex. 3.8 in Chap. 3.2.

Example Ex. 3.10 in Chap. 3.2.

2.1 Parameter of a distribution

For example, a Bernoulli variable takes value 0 with probability a and 1 with probability $1 - a$. a is a parameter. When a takes any specific value in the set of its permissible values (here $[0, 1]$), we get a specific distribution (a Bernoulli distribution). We say all the distributions with a taking different values constitute a distribution family (here, the Bernoulli family).

Example Ex. 3.12 and 3.14 in Chap. 3.2.

3 Cumulative distribution function (cdf)

Definition (for both discrete and continuous rv's):

$$F(x) = P(X \leq x)$$

In the discrete case: box on p. 97 (7th ed) or p. 104 (8th ed).

3.1 Obtain cdf from pmf

$$F(x) = \sum_{t:\, t \leq x} p(t)$$

3.2 Obtain pmf from cdf

$$p(x) = F(x) - F(x-)$$

where we use $x-$ to indicate the possible value of X that is "immediately" smaller than x.

3.3 Obtain probabilities from cdf

Using the same notation, for any $a \le b$:

$$P(a \le X \le b) = F(b) - F(a-)$$

In particular, if X takes integer values, then for integers $a \le b$:

$$P(a \le X \le b) = F(b) - F(a-1)$$

Exercise How do you obtain probabilities from pmf?

Example Ex. 3.13 in Chap. 3.2.

4 Mean and variance

We learn two basic properties, or characteristics, of a distribution: mean and variance. Mean describes the "center" of the distribution; variance describes the "spread", or variability, or dispersion, of the random variable around its mean.

Both mean and variance are the "expected values" of something. They are collectively called the first two "moments" of the distribution.

4.1 The expected value of a function of X

Since the value of a rv X is determined by chance, if we "draw" many samples of X (meaning, behind the scene, perform the experiment many times), it will take various values, hence any function of X, e.g. X^2 or $\sin X + 3$, will also take various values. The average value of this function will converge to a stable value if the sample size tends to infinity. The "limiting average" is called the "expected value" or "expectation" or "mean".

Let $h(X)$ be a function of X, then the <u>expected value</u> of $h(X)$ is denoted by $E\big[h(X)\big]$. Since we typically use μ to denote the expected value, this can also be written as $\mu_{h(X)}$ (note that $h(X)$ is in the subscript).

In the discrete case,

$$E\big[h(X)\big] = \sum_{\text{all possible value } x} h(x) \cdot p(x) \tag{1}$$

This is the weighted average of $h(x)$, using the pmf as the weight.

> One may be interested in any function of X, e.g. $h(X) = \log(X)$, $h(X) = X^2$, $h(X) = e^X + 3$, and so on.

4.2 Linearity of expected values

If $g(X)$ and $h(X)$ are any functions of X, and a, b, c are constants, then the expected value of a linear function of $g(X)$ and $h(X)$, say $ag(X) + bh(X) + c$, is

$$E\big[ag(X) + bh(X) + c\big] = aE\big[g(X)\big] + bE\big[h(X)\big] + c$$

Proof: use definition (1).

We say the "expectation" function $E(\cdot)$ is linear.

A special case is the following relation:

$$E[aX + b] = aE[X] + b.$$

4.3 Mean

The expected value or expectation or mean of the random variable X itself is denoted by $E(X)$ (or $E[X]$) or μ or μ_X.

Its definition is obtained by taking $h(X)$ to be X in (1), that is,

$$E[X] = \sum_{\text{all } x} x \cdot p(x)$$

This is the most commonly used measure of the location or center of a distribution. Other measures of center include median, etc.

Example Ex. 3.18 in Chap. 3.3. (Bernoulli)

> About math symbols or notation: one may choose one's own notation as long as the usage is clear and consistent. However, common usage should be followed whenever possible, to minimize confusion and ease understanding. For example, π. Using μ to mean "population mean" (not sample mean) is

pretty established, although not as unshakable as π.

Note Distinguish between population mean $E[X]$ and sample mean $\overline{X}$. Population mean is known only if the distribution is known. Sample mean is the simple arithmetic average calculated for a given data set (i.e. sample).

Caution $E(X)$ does not need to be a permissible value for X. This happens especially often with discrete random variables. For example, if X takes integer values, $E(X)$ often is not an integer.

4.4 Variance

Take $h(X) = (X - \mu)^2$, then $E\big[h(X)\big]$ is called the variance, denoted by $\mathrm{var}(X)$ or σ^2.

In the discrete case, using definition (1) one obtains

$$\mathrm{var}(X) \equiv E\big[(X-\mu)^2\big] = \sum_{\text{all possible } x} (x-\mu)^2 \cdot p(x) \quad (2)$$

Variance is the expected value of the squared deviation of X from the mean. It measures the "spread", or "variability", or "dispersion" of X about its mean. This is the most commonly used measure of spread of a distribution. Other useful measures of variability include "mean absolute deviation (mad)", $E\big[|X-\mu|\big]$, and "inter-quartile range" (IQR), $Q_3 - Q_1$.

The (positive) square root of variance is called standard deviation, denoted by σ. $\sigma = \sqrt{\mathrm{var}(X)}$.

The unit of σ^2 is that of X squared. The unit of σ is that of X.

Note Distinguish population variance σ^2 from the sample variance s^2.

4.4.1 Non-linearity of variance

What is the variance of a linear function of X? In general, we have

$$\mathrm{var}(aX+b) = a^2\,\mathrm{var}(X)$$

Proof

$$
\begin{aligned}
\operatorname{var}(aX+b) &= E\big[(aX+b)-E(aX+b)\big]^2 && \text{(definition of variance)}\\
&= E\big[(aX+b)-aE(X)-b\big] && \text{(property of } E(\cdot))\\
&= E(aX-a\mu)^2\\
&= a^2E(X-\mu)^2 && \text{(property of } E(\cdot))\\
&= a^2\operatorname{var}(X) && \text{(definition of variance)}
\end{aligned}
$$

This indicates that the variance function $\operatorname{var}(\cdot)$ is nonlinear.

> In words, adding a constant does not change the variance (it just "shifts" the distribution but does not change its spread); multiplying a factor gets the variance multiplied by the factor squared.

Note: In the proof above, we did not use definitions specific for discrete or continuous cases, hence the result holds in both discrete and continuous cases.

More generally, we have

$$\operatorname{var}\big[ag(X)+b\big] = a^2\operatorname{var}\big[g(X)\big]$$

where $g(X)$ is a function of X, and a, b are constants.

4.4.2 A property useful for calculation

$$\operatorname{var}(X) = E(X^2) - [E(X)]^2 \qquad (3)$$

Proof

$$
\begin{aligned}
\operatorname{var}(X) &= E[(X-\mu)^2]\\
&= E[X^2-2\mu X+\mu^2]\\
&= E[X^2]-2\mu E[X]+\mu^2\\
&= E[X^2]-2\mu^2+\mu^2
\end{aligned}
$$

> Formula (3) saves us from doing the subtractions in the definition (2).

Note: In the proof above, we did not use the particular forms of definition for the discrete or continuous case, hence the result hold in both discrete and continuous cases.

Exercise: Prove the relation above for the discrete case using the definition (2).

Example: Let X = the number of cylinders in the engine of the next car to be tuned up at a certain facility. The distribution of X is given by its pmf:

x	4	6	8
$p(x)$	.5	.3	.2

1. Calculate $E[X]$.
2. The cost of a tune-up is related to X by $h(X) = 20 + 3X + .5X^2$. Calculate $E[h(X)]$.
3. Calculate $\mathrm{var}[X]$ by definition (2).
4. Calculate $\mathrm{var}[X]$ by formula (3).

5 Bernoulli distribution

(This is the simplest discrete distribution. And it's useful.)

Example Toss an unbalanced coin, assume $P(\{H\}) = 0.48$ and $P(\{T\}) = 0.52$. Assign numerical values 0 and 1 to head and tail, respectively, then we have a discrete (binary, to be specific) random variable. Its distribution is

$$p(x) = \begin{cases} 0.48, & x = 0 \\ 0.52, & x = 1 \end{cases}$$

In general, if $p(1) = r$, where $0 < r < 1$, we can write

$$p(x) = \begin{cases} 1 - r, & x = 0 \\ r, & x = 1 \end{cases}$$

or more concisely,

$$p(x) = r^x(1-r)^{1-x}, \quad x = 0, 1$$

This is called the Bernoulli distribution.

Example The mean and variance of a Bernoulli variable.

Using the definitions of mean and variance, we get

$$E[X] = 1 \times r + 0 \times (1 - r) = r$$

and

$$\mathrm{var}[X] = (1-r)^2 r + (0-r)^2(1-r) = r(1-r)$$

6 Binomial distribution

6.1 Definition

6.1.1 Binomial experiment

Satisfies all the following requirements

1. Consists of n sub-experiments, called "trials". n is fixed prior to the experiment.
2. Each trial can result in one of two possible outcomes, referred to as S (success) and F (failure).

   The notation S and F are just for the purpose of distinguishing two outcomes. It does not need to mean S is good, F is bad, and so on.

3. Trials are independent. The outcome of one trial is not affected by the other trials.
4. The probability of S is constant from trial to trial. This probability is denoted by r. (Often denoted by p; but we want to avoid confusion with the pmf symbol p.)

6.1.2 Binomial random variable

In a binomial experiment, the focus of interest is usually how many S have resulted in the n trials. This count of S is called a "binomial random variable", X, denoted by

$$X \sim \text{Binom}(n, r)$$

Note: we do not care which trials are S and which ones are F; we only care about the total number of S.

Example The same coin is tossed successively and independently n times. Call head a S and tail a F; suppose $r = .5$.

6.2 The distribution

Following the textbook, we use $b(x; n, r)$ and $B(x; n, r)$ to denote binomial pmf and cdf, respectively.

First, the possible values of X are $0, 1, \ldots, n$.

Second, to calculate the probability $P(X = x)$, notice the event contains $\binom{n}{x}$ mutually exclusive sub-events (situations), and the probability of each of them is $r^x(1-r)^{n-x}$. Hence

$$b(x; n, r) = \binom{n}{x} r^x (1-r)^{n-x} \quad x = 0, 1, 2, \ldots, n$$

$$B(x; n, r) \equiv P(X \leq x) = \sum_{y=0}^{x} b(y; n, r), \quad x = 0, 1, \ldots, n$$

Computing $B(x; n, r)$ by its definition (see above) is almost always too tedious. There are three alternatives: (1) look up pre-made tables; (2) use a statistical computer package; (3) use normal approximation. This material is not required.

6.3 Mean and variance

If $X \sim \text{Binom}(n, r)$, then

$$E[X] = nr, \quad \text{var}[X] = nr(1-r)$$

Proof

The $\text{Binom}(n, r)$ variable X can be written as $X = Y_1 + Y_2 + \cdots + Y_n$, where Y_i are independent Bernoulli random variables with success rate r. We know $E[Y_1] = \cdots = E[Y_n] = r$ and $\text{var}[Y_1] = \cdots = \text{var}[Y_n] = r(1-r)$.

Using properties of expectation and variance, we have

$$E[X] = E[Y_1] + E[Y_2] + \cdots + E[Y_n] = nr$$

$$\text{var}[X] = \text{var}[Y_1] + \text{var}[Y_2] + \cdots + \text{var}[Y_n] = nr(1-r)$$

These two properties of the E and var operators will be introduced later. They hold for both discrete and continuous distributions. The first property does not require the Y_i's to be independent. The second does.

6.4 A typical approximate binomial experiment

Suppose there are N identical balls except for color: k of the balls are red and the others are blue. Thoroughly mix the balls. Randomly pick n balls and denote the number of red ones picked by X.

Rule of thumb: if $n \leq .05N$ and r is not too close to 0 and 1, then the experiment can be considered a binomial experiment and X is distributed approximately as $\text{Binom}(X; n, r)$, where $r = k/N$.

An alternative statement of the condition is $n \ll Nr$ and $n \ll N(1-r)$.

Example Ex. 3.30 in Chap. 3.4.

Example Polls. Suppose Gallup polls Americans on a certain policy, and proportion r of population are for it and $1 - r$ against. Suppose 2000 Americans are randomly chosen from the population such that for any person surveyed, the chance that this is a supporter is r. Now this is sampling without replacement, but the sample size is much smaller than the population.

7 Useful `R` functions

`choose`, `factorial`, `combn`

`sample` (take a look at its argument `prob`)

`rbinom`, `dbinom`, `pbinom`, `qbinom`